TEACHER PREPARATION PROGRAMS AND TEACHER QUALITY: ARE THERE REAL DIFFERENCES ACROSS PROGRAMS?

Cory Koedel
(corresponding author)
Economics Department
University of Missouri
Columbia, MO 65211
koedelc@missouri.edu

Eric Parsons
Economics Department
University of Missouri
Columbia, MO 65211
parsonses@missouri.edu

Michael Podgursky
Economics Department
University of Missouri
Columbia, MO 65211
podgurskym@missouri.edu

Mark Ehlert
Economics Department
University of Missouri
Columbia, MO 65211
ehlertm@missouri.edu

Abstract

We compare teacher preparation programs in Missouri based on the effectiveness of their graduates in the classroom. The differences in effectiveness between teachers from different preparation programs are much smaller than has been suggested in previous work. In fact, virtually all of the variation in teacher effectiveness comes from within-program differences between teachers. Prior research has overstated differences in teacher performance across preparation programs by failing to properly account for teacher sampling.

doi:10.1162/EDFP_a_00172

1. INTRODUCTION

There is increased national interest in holding teacher preparation programs (TPPs) accountable for how their graduates perform in the classroom. In a 2010 report from the Center for American Progress, Crowe (2010, p. 2) argues that "every state's teacher preparation program accountability system should include a teacher effectiveness measure that reports the extent to which program graduates help their K–12 students to learn." This sentiment is echoed by Aldeman et al. (2011) and the U.S. Department of Education (USDOE 2011). Numerous states have or are building the capacity to evaluate TPPs using longitudinal state data systems that link students to their teachers. In fact, all twelve phase-1 and phase-2 winners of the federal Race to the Top competition have committed to using student achievement outcomes for TPP evaluations, and five will use estimates of teacher impacts on student achievement for program accountability (Crowe 2011). Some states—notably Louisiana and Tennessee—have for several years been reporting estimates that associate TPPs with student-achievement growth. The Louisiana model in particular has received considerable national attention. Paul G. Pastorek, the former state superintendent in Louisiana, shares his sentiment regarding the Louisiana model in a report released by the U.S. Department of Education in 2011: "I applaud the U.S. Department of Education for working to take the Louisiana-model nationwide. Teacher preparation program accountability for K–12 results is an idea whose time has come" (USDOE 2011, p. 7).

This paper offers a sobering view of TPP accountability of this form, at least in current application. We use a statewide longitudinal data set from Missouri—similar to other data sets that have been used for TPP evaluations elsewhere—to examine the extent to which teachers who are prepared by different TPPs differ in effectiveness. We measure the effectiveness of teachers from different preparation programs using commonly applied empirical models. The key finding from our analysis is that teachers from different training programs differ much less than has been indicated by prior research in terms of their ability to raise student achievement. This suggests that TPP-of-origin is a less useful indicator for educational administrators looking to hire effective teachers than has been implied by earlier work.

In TPP evaluations, it is important to recognize that students who are taught by the same teacher are not independent observations regarding the effectiveness of that teacher's preparation program. The level of clustering of the standard errors in models of TPP effects can significantly influence the interpretation of results. Although one could argue that multiple levels of data clustering are important in models that evaluate TPPs, prior research

suggests the most important level is that of individual teachers.[1] Previous studies have not clustered at the teacher level and in doing so have reported standard errors that are too small (e.g., Noell, Porter, and Patt 2007; Noell et al. 2008; Gansle et al. 2010; Gansle, Noell, and Burns 2012). As a result, they have misinterpreted a portion of the sampling variability in the data to represent real differences in teaching effectiveness across teachers from different TPPs.[2] After clustering at the teacher level, we find that most, if not all, of the variation in the estimated TPP effects in Missouri can be attributed to estimation-error variance.

Our finding that true TPP effects are much smaller than has been previously indicated is robust to different specifications for the student-achievement model and persists despite the fact that we observe, on average, over fifty teachers per training program. This is more than the average program in the Louisiana and Tennessee evaluations (Gansle et al. 2010; Tennessee Higher Education Commission 2010).[3] Furthermore, our findings hold in a subsample of large TPPs for which we observe an average of more than eighty teachers per program. The implication is that our results are not driven by particularly weak data conditions in Missouri—other states attempting to perform TPP evaluations will likely have similar or inferior data (at least relative to our large-program subsample), particularly in instances where broad inclusion of TPPs is desired for accountability purposes (e.g., see Lincove et al. 2013).

The lack of true variation in TPP effects uncovered by our study is perhaps surprising because researchers have documented what appear to be considerable input-based differences across TPPs (along dimensions such as course offerings, program emphasis, mentoring, etc. For example, see Levine 2006; Boyd et al. 2009). Given the importance of differences in effectiveness between individual teachers in determining student outcomes (Hanushek and Rivkin 2010) and the apparent variation in program inputs across TPPs, it would seem reasonable to hypothesize that differences in how teachers are prepared across programs would translate into differences in how they perform in the classroom. We find no evidence to support this hypothesis, however. One explanation is that the input-based differences in how teachers are trained across

1. It has been well established that there are large differences in effectiveness across individual teachers. See Hanushek and Rivkin (2010) for a review of the recent literature.
2. In the Gansle et al. (2010), Noell, Porter, and Patt (2007), and Noell et al. (2008) studies, clustering occurs at the school and classroom levels, where teachers can teach in multiple classrooms. The classroom-level clustering is helpful but will still overstate statistical power, particularly as the ratio of classrooms to teachers increases in the data. The comparison between classroom and teacher-level clustering in this context is akin to the comparison between clustering at the state and state-by-year levels in Bertrand, Duflo, and Mullainathan (2004). We discuss the clustering issue in more detail in section 5.
3. Our data also include more programs with fifty or more linked teachers than in these other evaluations.

TPPs are not as large as they appear to be—that is, it may be that most TPPs are providing similar training.[4]

Our study makes two substantive contributions to the literature on TPP evaluation. First, we recommend a technical correction to models that are used to evaluate TPPs—teacher-level clustering—and show that with improper clustering the models produce misleading results. Second, after making the technical correction, we show that differences in effectiveness across graduates from different TPPs are much smaller than has been previously suggested. Virtually all of the variation in teacher effectiveness in the labor force occurs across teachers within programs. We conclude, therefore, that TPP rankings based on growth model output are, at present, of little value to state departments of education, TPP accreditation agencies, and K–12 school administrators. If it is not made clear that the substantive differences that separate TPPs in the rankings are small, the rankings could lead to inefficient hiring and other decisions.

Several qualifications to our findings are in order. First, our analysis focuses entirely on teacher evaluations based on student achievement. Although achievement-based evaluations have become a focal point for recent policy initiatives (Crowe 2011), they represent just one measure of the effectiveness of TPP graduates.[5] Second, our evaluation includes only traditional TPPs—there may be additional heterogeneity across programs in analyses that also consider nontraditional TPPs (e.g., alternative-certification programs), in which case real differences in effectiveness across programs may emerge.[6] Third, our findings cannot speak to whether continued efforts to evaluate and rank TPPs based on how teachers perform in the classroom will be fruitful. For example, a recent study by The New Teacher Project (Weisberg et al. 2009)

4. Regardless of the cause, our findings are consistent with TPP-of-attendance being another intervention that does not influence teaching effectiveness (e.g., see Harris and Sass 2011; for a recent exception see Taylor and Tyler 2012). In Appendix B, we also show that differences across TPPs in student selectivity are smaller than what is implied by the differences in selectivity between the typical students from the institutions that house the TPPs. Further, although the TPP effectiveness measures estimated in this paper combine selection and input-based effects (see section 3 for further discussion), as long as these effects are not offsetting we might still expect the input-based differences to show up in our TPP effectiveness measures.
5. There is a growing literature that examines the use of composite measures of teaching effectiveness along multiple dimensions (e.g., see Mihaly et al. 2013b; Bill and Melinda Gates Foundation 2013). Future TPP evaluations could be expanded to evaluate teachers using composite measures. Nevertheless, a key finding from this study is that small TPP-level sample sizes, combined with small differences in teacher quality across teachers from different TPPs, create a fundamental power problem with these types of evaluations; using composite measures may do little to resolve the power problem.
6. Also, of course, we cannot rule out that TPP of attendance may be a more important predictor of teacher performance in other states, even among traditional TPPs. It is noteworthy, however, that the reports from Louisiana (Noell, Porter, and Patt 2007; Noell et al. 2008; Gansle et al. 2010) and Tennessee (Tennessee Higher Education Commission 2010) show what we interpret to be small substantive differences between teachers from different TPPs.

suggests that the general lack of accountability within the education sector has led to complacency.[7] Indeed, our non-findings may reflect the fact that TPPs have had little incentive thus far to innovate and improve. The mere presence of annual rankings, even if they are initially uninformative, may prompt improvements in teacher preparation moving forward. Even small improvements could greatly improve students' short-term and long-term outcomes (Hanushek and Rivkin 2010; Chetty, Friedman, and Rockoff 2014; Hanushek 2011).

2. DATA

We use statewide administrative data from Missouri to evaluate recent TPP graduates from traditional, university-based programs. We begin with the universe of active teachers in elementary classrooms in Missouri during the 2008–09 school year, which is the first year for which we have linked student–teacher data. From these teachers, we identify the subset who began teaching no earlier than the 2004–05 school year. We then follow them for up to two additional years beyond the 2008–09 school year, through 2010–11; this allows us to observe up to three classrooms per teacher.

We link teachers to their certification records as provided by the Department of Elementary and Secondary Education and consider all teachers who were recommended for certification by a major Missouri institution within three years of their date of first employment (consistent with the policy focus on the effectiveness of recent graduates). For the purposes of our analysis we define a "major" Missouri institution liberally—we require the institution to have produced more than fifteen active teachers in our data set. We also separately evaluate the subset of TPPs that produced more than fifty teachers in our analytic sample.[8]

Students in Missouri are first tested using the Missouri Assessment Program exam in grade 3, which means that grade 4 teachers are the first teachers for whom we can estimate value added to student scores. Therefore, our analysis includes all teachers in self-contained classrooms in elementary schools

7. For example, The New Teacher Project report finds that, for most teachers, areas of improvement are not identified on their annual evaluations.
8. A general issue in TPP evaluation is within-institution heterogeneity across programs. For example, a single institution may offer multiple certification programs across schooling levels and subjects and/or alternative certification routes. Our focus on traditional TPP programs and on teachers moving into elementary schools (a relatively homogenous output sample) reduces within-institution heterogeneity. In our primary results, we simply compare all of the teachers from each program who end up in self-contained elementary classrooms in Missouri public schools. Further analysis reveals that just 1.4 percent of our main sample is identified in the certification files as obtaining an alternative certification from one of the TPPs that we evaluate. Given this low number, it is unsurprising that our findings are unaffected by our decision of whether or not to include these teachers in the analytic sample (nonetheless, results from models where alternatively certified teachers are omitted from the analytic sample are available upon request).

in grades 4, 5, and 6.[9] Our final sample includes 1,309 unique teachers who were certified from one of the 24 major preparation programs in the state. These teachers are spread across 656 elementary schools.

The teachers in our sample can be linked to 61,150 student-year records with current and lagged math test scores, and 61,039 student-year records with current and lagged communication arts scores. The student-level data include basic information about race, gender, free/reduced-price lunch status, language-learner status, and mobility status (whether the student moved schools during the course of the school year). We construct school-level aggregates for each of these variables as well, which we include in some of our models. Table 1 provides basic summary information for the data, and table A.1 in Appendix A lists the teacher counts from the 24 preparation programs (column 1 shows teacher counts for the main analysis). To maintain the anonymity of the TPPs we use generic program labels throughout our study.[10]

3. EMPIRICAL STRATEGY

Estimation of TPP Effects

We follow the approach used by several other recent TPP studies and estimate empirical models of the following form (Boyd et al. 2009; Goldhaber, Liddle, and Theobald 2013; Mihaly et al. 2013a):

$$Y_{ijsgt} = Y_{ijsg(t-1)}\delta_1 + X_{ijsgt}\delta_2 + S_{ijsgt}\delta_3 + T_{ijsgt}\delta_4 + TPP_{ijsgt}\theta + \phi_g + \pi_t + \varepsilon_{ijsgt}. \quad (1)$$

In equation 1, Y_{ijsgt} is a test score for student i taught by teacher j at school s in grade g and year t, standardized within the grade-subject-year cell. X_{ijsgt} includes basic demographic and socioeconomic information for student i; S_{ijsgt} includes similar information for the school attended by student i; T_{ijsgt} includes controls for teacher experience; and TPP_{ijsgt} is a vector of TPP indicator variables where the entry is set to one for the program from which teacher j, who teaches student i, was certified.[11] ϕ_g and π_t are grade and year fixed effects. We perform

9. Our grade 6 sample is only a partial sample of grade 6 teachers in the state, as in many districts grade 6 is taught in middle schools and therefore there are no self-contained grade 6 teachers. Note that we generally cannot observe departmentalized teaching within elementary schools in Missouri—to the extent that elementary students' classroom teachers are not teaching math and/or communication arts to their students, our estimates may be attenuated.
10. This is at the request of the Missouri Department of Elementary and Secondary Education.
11. In unreported results, we verify that including additional controls for classroom characteristics does not affect our findings qualitatively. We also obtain similar results if we use a model analogous to equation 1 to estimate individual teacher effects and then produce TPP effect estimates by aggregating the estimated teacher effects. A conceptual difference between the method we use in the text and the "aggregated teacher effect" method is in the identifying variation that is used to estimate the coefficients on the variables in the X-, S-, and T-vectors. Our preferred approach, shown in

Table 1. Data Details

Primary Data Set	
Preparation programs evaluated	24
New teachers in grades 4, 5, and 6 who were verified to receive a certification and/or degree from a valid institution within three years of start date[a]	1,309
Maximum number of teachers from a single preparation program	143
Minimum number of teachers from a single preparation program	16
Number of schools where teachers are observed teaching	656
Number of student-year records with math test scores that could be linked to teachers	61,150
Number of student-year records with communication arts test scores that could be linked to teachers	61,039
Average number of classrooms per teacher	2.38
Programs Producing 50 or More New Teachers[b]	
Preparation programs evaluated	12
New teachers in grades 4, 5, and 6 who were verified to receive a certification and/or degree from a valid institution within three years of start date[a]	1,000
Maximum number of teachers from a single preparation program	143
Minimum number of teachers from a single preparation program	49
Number of schools where teachers are observed teaching	555
Number of student-year records with math test scores that could be linked to teachers	46,702
Number of student-year records with communication arts test scores that could be linked to teachers	46,628
Average number of classrooms per teacher	2.38

Notes: [a]We only include teachers who were in self-contained classrooms in these grades (e.g., elementary schools). Many grade 6 teachers teach in middle schools in Missouri.
[b]Program 12 was also included in this group although it had only 49 new teachers represented in the sample (see Appendix table A.1).

our analysis separately for student achievement in math and communication arts.[12]

equation 1, uses within-TPP variation. The aggregated-teacher-effect approach uses within-teacher variation. A detailed discussion of the strengths and weaknesses of these two approaches is beyond the scope of the present paper—we refer the interested reader to Ehlert et al. (forthcoming). The important substantive point for the present study is that our findings are qualitatively unaffected by this and other modeling decisions that we have considered (also see section 6).

12. Other notable studies—Noell, Porter, and Patt (2007), Noell et al. (2008), and Gansle et al. (2010)—use a multilevel model that differs mechanically from the model used here but is very similar conceptually. We consider similar models in section 6. Goldhaber, Liddle, and Theobald (2013) also extend the general framework to account for the decay of TPP effects over time. Decay in the TPP effects is one potential explanation for why they are so small, particularly when TPP effects are estimated using data from multiple cohorts of teachers. All of the studies of which we are aware evaluate TPPs using multiple cohorts to increase teacher sample sizes. Our analysis indicates that the sample-size issue is even more important than is implied by prior studies.

Notice that the model does not include indicators for individual teachers. We do not condition on individual teacher effects simultaneously with TPP effects in the model because the objective is to attribute teacher performance to the TPPs. Nevertheless, a large body of research shows that there are considerable differences in effectiveness across individual teachers that persist across classrooms and over time (e.g., see Goldhaber and Hansen 2010; Hanushek and Rivkin 2010; Goldhaber and Theobald 2013). These differences create a clustering structure within the data. For example, if students *A* and *B* are both taught by teacher *Q* who was trained at program *Z*, the two students cannot be treated as independent observations by which the effectiveness of teachers from program *Z* can be identified. Our standard errors are clustered at the individual-teacher level throughout our analysis to reflect this data structure.

We consider models that include several teacher characteristics, but in our primary analysis we control only for teacher experience as shown in equation 1. Research consistently shows that teacher performance improves with experience. Because we evaluate five different cohorts of entering teachers beginning in a single year (2008–09), the experience control is important so that differences in the experience profiles of teachers across training programs are not confounded with the program impacts.[13] Goldhaber, Liddle, and Theobald (2013) show that whether the model includes controls for other observable teacher characteristics is of little practical consequence for evaluating TPPs. This is the case in our data as well (results not shown for brevity).[14]

We estimate the model in equation 1 with and without school characteristics and present our findings from each specification in math and communication arts. We present models that compare all 24 programs and models that compare the 12 "large" programs (those that produced more than 50 teachers in our data). We test for the joint significance of the program indicators using F-tests.

An additional consideration is whether equation 1 should be estimated with school fixed effects. A benefit of the school-fixed-effects approach is that it removes any bias owing to systematic differences across TPPs in the quality of the K–12 schools where graduates are placed. To identify the relative program impacts, however, it also relies on comparisons among teachers at K–12 schools that house graduates from multiple preparation programs. That is,

13. It is uncontroversial that performance improves with experience for teachers in the early years of their careers (many studies are available; see, for example, Clotfelter, Ladd, and Vigdor 2006). Recent studies by Wiswall (2013) and Papay and Kraft (2010) find that experience matters further into teachers' careers.

14. Although we control for teacher experience in our models, there is a potential selectivity concern that the models do not account for directly. Namely, if ineffective teachers are disproportionately trained by some TPPs and also have different attrition patterns, then this could influence our comparative findings. We examine this possibility in section 6.

TPP estimates from school-fixed-effects models depend on teachers who teach in K–12 schools where teachers from other TPPs are also teaching. The K–12 schools that house teachers from multiple programs, and the teachers who teach at these schools, may or may not be useful for gaining inference about the larger program effects. For example, it may be the case that teachers who select into teaching at the same school are more similar in their effectiveness than teachers who teach at different schools, in which case TPP effect estimates from school-fixed-effects models will suffer from attenuation bias. Mihaly et al. (2013a) provide a thorough analysis of the tradeoffs involved in moving to a school-fixed-effects specification. Rather than replicate their discussion here, we simply note that these tradeoffs exist. Our key result—that differences in teacher effectiveness across TPPs are much smaller than has been previously suggested—is obtained regardless of how we set up the model. We do not directly report here estimates from models that include school fixed effects but these estimates are available from the authors upon request.

Analysis of TPP Effects

After estimating the model in equation 1 for math and communication arts achievement, we extract the estimated TPP effects. A key policy question is this: How much do graduates from different TPPs differ in effectiveness? It is important to recognize that this is not the same question as "what is the efficacy of the training provided by different TPPs?" The latter question aims to identify TPP training effects free from the effects of initial selection into the programs. Separating out selection effects is unlikely to be of great interest to administrators in the field, however. For a district administrator, the question of *why* teachers from one TPP outperform teachers from another is not nearly as important as simply identifying the programs that, on the whole, graduate the most effective teachers. Indeed, in all of the locales where achievement-based metrics are being used to evaluate TPPs, selection effects and training efficacy are wrapped into a single estimate. We proceed with the primary objective of estimating this combined effect, consistent with current policy practice.[15]

We produce several measures of the variability in teacher effectiveness across TPPs. The first measure is the increase in the overall (unadjusted) R^2 in the model when we add the TPP indicators. The predictive power of the

15. Even if we could separate out selection effects from TPP training effectiveness, it may still be desirable to evaluate TPPs based on the combined effect. For example, we may want to reward TPPs that are successful in bringing talented individuals into the teaching profession. On the other hand, in terms of developing a more effective training curriculum for teachers, understanding TPP training effectiveness is of primary interest. Data that provide more detail about the experiences of individual teachers within TPPs, similar to the data used by Boyd et al. (2009), would be particularly valuable for learning more about which aspects of training are most important.

TPP indicators reflects systematic differences in teacher effectiveness across graduates from different programs. If the programs do not differ, we would expect the change in R^2 to be zero when the program indicators are included. We compare the change in R^2 from adding the TPP indicators to the change in R^2 from adding individual teacher indicators in their place. The change in R^2 when we add the individual teacher indicators provides a measure of the total variability in teaching effectiveness across the teachers in our data sample. The ratio of the explanatory power of the TPP indicators to the explanatory power of the individual-teacher indicators provides a measure of the share of the total variance in teacher quality that can be explained by cross-program differences.

We also estimate the variance and range of program effects. Both measures are prone to overstatement because the TPP effects are statistical estimates; even in the absence of any real TPP effects, we would expect to estimate a nonzero variance and range of the TPP coefficients. Take the range—the value of the largest point estimate minus the smallest point estimate—as an example. Noting that each TPP coefficient is a composite of the true effect plus error, $\hat{\theta}_j = \theta_j + \lambda_j$, the range is determined partly by the estimation-error component. As the share of the total variance in the TPP coefficients attributable to estimation error rises, so does the overstatement of the estimated range.

Following the recent empirical literature on teacher quality, we decompose the variance in the TPP effects into two components: the true variance share and the estimation-error variance share.

$$Var(\hat{\theta}) = Var(\theta) + Var(\lambda). \tag{2}$$

In equation 2, $Var(\theta)$ is the true variance in the program effects, $Var(\hat{\theta})$ is the variance of the estimates from equation 1, and $Var(\lambda)$ is the estimation-error share of the variance. $Var(\hat{\theta})$ is estimated by the raw variance of the TPP coefficients. We provide an estimate of the variance due to true variability in the TPP effects, separated from estimation-error variance, using an ad hoc procedure based on Koedel (2009) and Mas and Moretti (2009). Specifically, we estimate $Var(\theta)$ as $\{Var(\hat{\theta}) - \frac{Var(\hat{\theta})}{Q}\}$. Q is a ratio where the numerator is the F-statistic from a test of the null hypothesis of joint equality for the TPP coefficients and the denominator is the 5-percent critical value. If the F-statistic is less than the critical value, indicating the variability in the TPP indicators cannot be statistically distinguished, $Var(\theta) = 0$. As the F-statistic increases beyond the critical value, indicating more true variability across TPPs, Q increases and correspondingly the estimate of $Var(\theta)$ rises. $Var(\theta)$ approaches $Var(\hat{\theta})$ as the F-statistic approaches infinity.[16]

16. We perform the calculations in table 3 using TPP estimates from models where we mean-center all of the data, include indicators for all TPPs, and omit the intercept term. We also replicate our

Table 2. Correlation Matrix for TPP Effects Estimated from Different Models

	Math A	Math B	Comm Arts A	Comm Arts B
Math A	1.00			
Math B	0.98	1.00		
Com-Arts A	0.63	0.57	1.00	
Com-Arts B	0.65	0.65	0.95	1.00
Student Covariates	X	X	X	X
School Covariates		X		X

Notes: Models A and B are as described in the text. Correlations are based on the models that include all 24 TPPs. Com-Arts: communication arts.

We use this approach to approximate the share of the total variance in the estimated TPP effects that reflects actual differences in program quality, and we report estimates of the adjusted standard deviation and range of the TPP effects. The true range of TPP effects will always be smaller than the observed range of point estimates; for some observed range Z, the expected value of the true range is $\sqrt{\frac{\text{var}(\theta)}{\text{var}(\hat{\theta})}} * Z$ (note that the true range can depend on a different pair of TPPs than the point-estimate range).[17]

4. RESULTS

Table 2 shows correlations between the TPP effects from the different models in math and communication arts. We estimate two different models in each subject:

> Model A: Includes the lagged student test score, student-level controls, controls for teacher experience, and the preparation program indicators
>
> Model B: Includes everything in Model A, plus school-level aggregates analogous to the student-level controls

Table 2 shows that within subjects, the estimates from Models A and B are similar, with correlations at or above 0.95 in math and communication arts. This suggests that observable differences in the K–12 schooling environments for graduates from the different TPPs introduce little bias into the estimates from Model A. Across subjects and within models, the correlation in the TPP effects remains above 0.60.

results qualitatively using models where we omit a comparison program and estimate TPP effects relative to a hold-out program.

17. In section 6 we also consider an alternative approach to isolate the true variance in the TPP effects.

Table 3. Variation in Preparation Program Effects

	Mathematics		Communication Arts	
	Model A	**Model B**	**Model A**	**Model B**
24 Major Preparation Programs				
p-value of joint F-test on TPP effects	0.014	0.008	0.001	0.003
ΔR^2 from adding TPP indicators	0.001	0.001	0.001	0.001
ΔR^2 from adding teacher-level indicators	0.045	0.044	0.027	0.026
ΔR^2 Ratio	0.027	0.029	0.034	0.031
Unadjusted standard deviation of TPP effects	0.040	0.042	0.037	0.037
Unadjusted range of TPP effects	0.152	0.171	0.171	0.171
Estimation-error variance share	0.869	0.826	0.724	0.760
Adjusted standard deviation	0.015	0.018	0.019	0.018
Adjusted range	0.055	0.071	0.089	0.084
Programs Producing 50 or More New Teachers				
p-value of joint F-test on TPP effects	0.045	0.066	0.020	0.122
ΔR^2 from adding TPP indicators	0.001	0.001	0.001	0.000
ΔR^2 from adding teacher-level indicators	0.047	0.047	0.027	0.026
ΔR^2 Ratio	0.019	0.019	0.023	0.016
Unadjusted standard deviation of TPP effects	0.033	0.032	0.025	0.021
Unadjusted range of TPP effects	0.137	0.127	0.099	0.082
Estimation-error variance share	0.978	1.00	0.866	1.00
Adjusted standard deviation	0.005	0	0.009	0
Adjusted range	0.020	0	0.036	0

Notes: In cases where the *p*-value from the joint F-test on the TPP effects is greater than 0.05, indicating that there are no statistically significant differences across TPPs, a value of 1.00 is reported for the estimation-error variance share.

Table 3 reports estimates of the variance in effectiveness across teachers from different TPPs. We split the table into two parts. The first horizontal panel evaluates the 24 "main" TPPs in Missouri (i.e., with more than fifteen graduates in our data); the second horizontal panel evaluates just the "large" programs. The large program subsample includes only half of the programs but over 75 percent of the teachers in our data (see table 1). The large programs are diverse in terms of overall selectivity (based on all university entrants; see Appendix B), and our comparisons between these programs benefit from relatively large program-level teacher sample sizes. Specifically, the average number of teachers per large program is 83.3 (table 1).[18]

18. See Appendix table A.1 for program-by-program teacher counts.

We analyze the output from each model in the same fashion throughout the table. We begin by reporting the *p*-value from a simple F-test for the null hypothesis that the TPP indicators are jointly insignificant in the model. We reject the null hypothesis that the TPP indicators are equal to zero in all of the specifications where we evaluate the 24 TPPs (the top panel of the table). For the large-program models, we reject the null using Model A, but not Model B.

Although the *p*-values provide information about statistical significance, they are less informative about practical significance. We begin our investigation into the practical significance of the differences across TPPs by comparing the predictive power of the TPP indicators to the predictive power of individual-teacher indicators. The second row in each panel of table 3 shows the change in R^2 when the TPP indicators are added to the model, and the third row shows the change in R^2 when we add individual teacher indicators instead. The fourth row reports the ratio of these values, which we interpret as the share of the total variance in teacher performance that is explained by cross-program differences. The ratios reported in table 3 show that differences in teacher performance across teachers from different TPPs explain only a very small fraction of the total variance of the teacher effects. More specifically, cross-program differences explain no more than 3.4 percent of the total variance in teacher value added in any model, and as little as 1.6 percent. This provides the first indication that differences in teacher performance across TPPs are small—almost all of the variation in teacher value-added occurs within TPPs.

The next two rows in each panel of table 3 report the unadjusted standard deviation and range of the TPP effects from each model. The unadjusted standard deviation is calculated as the square root of the variance of the initial TPP estimates, and the unadjusted range is calculated by subtracting the smallest point estimate from the largest. In each of the models where we evaluate all 24 TPPs, the unadjusted standard deviation and range are large. For example, consider a standard deviation estimate of 0.04 across programs, which is close to what we report for the unadjusted standard deviation of the TPP effects in each model in the top panel of the table. If this estimate reflected true variability across programs, it would indicate a one-standard-deviation move across TPP effects is on the order of 20 to 40 percent of the size of a one-standard-deviation move in the distribution of individual-level teacher effects (Hanushek and Rivkin 2010). The unadjusted standard deviation and range are smaller, but still sizeable, when we focus on the large programs. However, neither of the unadjusted measures accounts for estimation error in the TPP effects.

The next three rows highlight the importance of accounting for estimation error. They show results after we make the estimation-error adjustment,

and reveal the overwhelming majority of the variation in the estimated TPP coefficients is the product of statistical noise—that is, most of the variance is unrelated to actual differences in TPP effects. In the top panel of the table the estimation-error adjustments indicate that the true variability in TPP effects is much smaller than what is implied by the point estimates from the model. In our comparisons involving the large TPPs this result is reinforced. Only a very small fraction of the variance in the TPP point estimates reflects true differences in teaching effectiveness across teachers from different TPPs.

Figure 1 illustrates the variability in the TPP effects from Model B for the large-program sample in math and communication arts. The estimates are shrunken following the approach used by Koedel, Leatherman, and Parsons (2012). Noting the scale of the vertical axis in the figure, all of the point estimates suggest modest relative effect sizes. Still, in both math and communication arts, some programs have confidence intervals that exclude zero. Comparing the figure to our results in table 3, it is clear that this can occur for some point estimates even when we cannot reject the null hypothesis that all program effects are equal.[19] Note that as the number of TPP effects being estimated increases, the number identified as statistically distinguishable from average by chance increases as well.

5. HOW CLUSTERING MATTERS

Earlier we noted that a key difference between our analysis and several prior studies is in the level of clustering.[20] It is important to recognize that any level of clustering below the teacher level overstates independence in the data, and therefore overstates statistical precision. For example, classroom clustering assumes students taught by the same teacher in different classrooms are independent observations. Teachers have persistent effects across classrooms, however, and because of this, students who are taught by the same

19. The *p*-values from the F-tests for the joint significance of the TPP effects presented in figure 1 are 0.066 for mathematics and 0.122 for communication arts (see table 3).

20. Gansle et al. (2010), Gansle, Noell, and Burns (2012), Noell, Porter, and Patt (2007), and Noell et al. (2008) cluster at the classroom level. Boyd et al. (2009) do not indicate a level of clustering in their study. In correspondence with the authors we were told that clustering occurs at the teacher level, although their standard errors are much smaller than the standard errors we report here despite their using a smaller estimation sample and models that include school fixed effects (which typically result in larger standard errors). We do not have an explanation for their findings. Finally, note that a conventional wisdom is that clustering should occur "at the level of the intervention," but clustering at the TPP level will result in unreliable standard errors in TPP evaluations because the number of data clusters is less than the number of parameters to be estimated (the parameters to be estimated are the coefficients for each TPP plus the coefficients for the other control variables in the model). Additionally, we question the rationale for clustering at the TPP level. There is little reason to expect that two students taught by two different teachers (perhaps at different schools) belong in the same cluster, particularly when one conditions on the TPP effects directly.

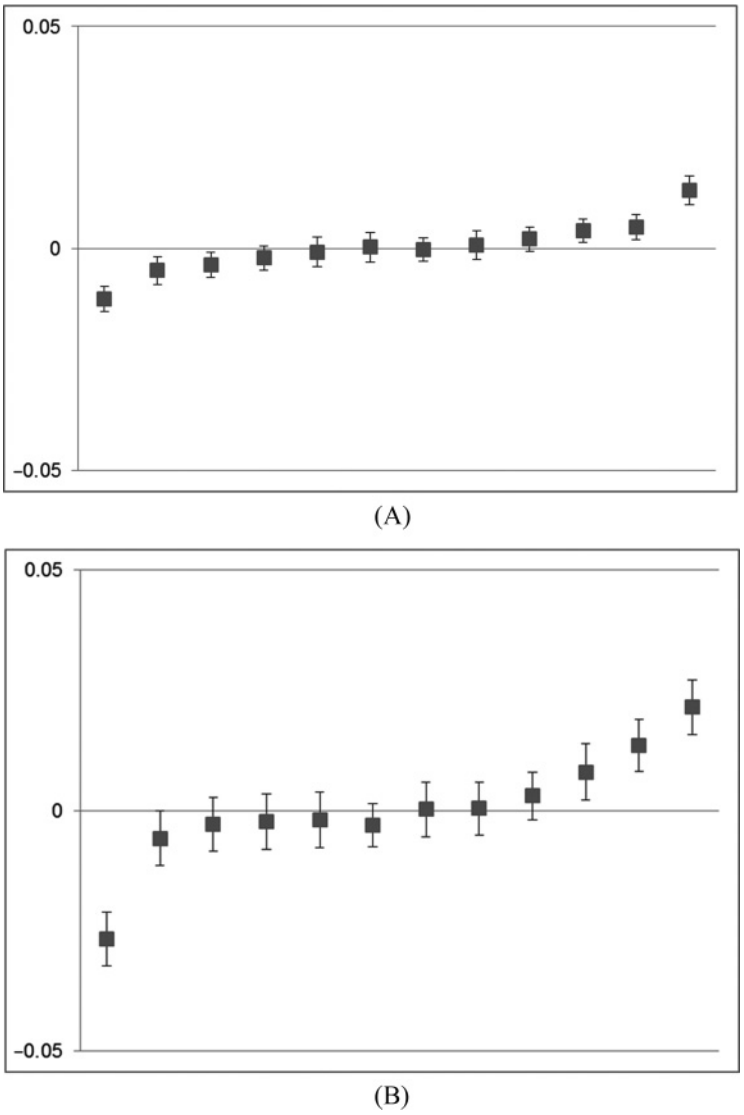

Figure 1. Distributions of the Shrunken TPP Effect Estimates from Model B, with Confidence Intervals: Programs Producing 50 or More New Teachers.
Panel A: Communication Arts
Panel B: Mathematics
Notes: The estimates in these figures are based on the output from Model B as shown in table 3. The estimates are shrunken ex post following Koedel, Leatherman, and Parsons (2012).

teacher—even across classrooms and/or years—cannot be viewed as independent observations regarding the effectiveness of that teacher's preparation program. The extent to which classroom clustering will overstate statistical precision depends on the persistence of teacher effects across classrooms and the ratio of classrooms to teachers in the data. As teachers are observed with

more and more classrooms, the overstatement of statistical power by models that cluster at the classroom level will increase.[21]

To illustrate the importance of clustering correctly, in table 4 we replicate our analysis from table 3 in math and communication arts without clustering the data, as well as with classroom-level clustering (we also show the teacher-level clustering results from table 3 for ease of comparison). Unsurprisingly, the models where we do not cluster at all imply differences across teachers from different TPPs that are substantially larger than what we report in table 3. For example, the *adjusted* standard deviation of the TPP effects from the large-program sample using math Model B, without clustering, is 0.029. Mapping this number back to the larger teacher quality literature would imply a one-standard-deviation move across TPP effects is on the order of 15 to 30 percent as large as a one-standard-deviation move in the distribution of individual teacher effects (Hanushek and Rivkin 2010). By contrast, with teacher-level clustering, the adjusted standard deviation across the large programs using Model B is zero.

Table 4 also shows that classroom-level clustering leads to estimation-error adjustments that are too small. Recall from section 2 that we observe up to three classrooms per teacher given the structure of our data, and on average we observe 2.4 classrooms per teacher in the analytic sample (see table 1). In analyses that span longer time horizons or where teachers teach multiple classes per year (e.g., middle or high school), the overstatement of statistical power from classroom clustering will be larger than what we show here.

Is the issue with TPP evaluation one of sample size? That is, if we could observe more teachers from each program could we statistically identify differences in TPP effects? Trivially, of course, it will be easier to detect small differences in TPP effects with larger sample sizes, but our sample sizes are not small for this type of analysis. Again, for the large programs in Missouri we observe an average of more than 80 teachers per program. This number is larger than what is reported in the 2010 reports from Louisiana (Gansle et al. 2010) and Tennessee (Tennessee Higher Education Commission 2010). Our sample sizes appear to be very reasonable, and even large, in terms of what can be expected from these types of evaluations.

21. Again, the comparison between classroom- and teacher-level clustering in this context is akin to the comparison between clustering at the state and state-by-year levels in Bertrand, Duflo, and Mullainathan (2004). In both intermediate cases (classroom-clustering in the present application or state-by-year clustering in the Bertrand, Duflo, and Mullainathan study), the clustering structure assumes too many independent observations, leading to standard errors that are too small. We also refer the interested reader to Bester, Conley, and Hansen (2011), who explore in detail the process by which researchers can choose cluster groups, including cases where clustering occurs along multiple dimensions.

Table 4. Replication of Results from Models A and B in Math and Communication Arts with Different Levels of Clustering

	Math Model A	Math Model A	Math Model A	Math Model B	Math Model B	Math Model B	Comm Arts Model A	Comm Arts Model A	Comm Arts Model A	Comm Arts Model B	Comm Arts Model B	Comm Arts Model B
Clustering	None	Clssrm	Tchr	None	Clssrm	Tchr	None	Clssrm	Tchr	None	Clssrm	Tchr
24 Major Preparation Programs												
p-value of joint F-test on TPP effects	0.000	0.000	0.014	0.000	0.000	0.008	0.000	0.000	0.001	0.000	0.000	0.003
Estimation-error variance share	0.173	0.594	0.869	0.172	0.574	0.826	0.241	0.500	0.724	0.259	0.528	0.760
Adjusted standard deviation	0.037	0.026	0.015	0.038	0.028	0.018	0.032	0.026	0.019	0.032	0.025	0.018
Adjusted range	0.138	0.097	0.055	0.156	0.112	0.071	0.148	0.121	0.089	0.147	0.117	0.084
Programs Producing 50 or More New Teachers												
p-value of joint F-test on TPP effects	0.000	0.001	0.045	0.000	0.002	0.066	0.000	0.001	0.020	0.000	0.017	0.122
Estimation-error variance share	0.170	0.623	0.978	0.182	0.665	1.00	0.303	0.616	0.866	0.418	0.850	1.00
Adjusted standard deviation	0.030	0.020	0.005	0.029	0.018	0	0.021	0.015	0.009	0.016	0.008	0
Adjusted range of TPP effects	0.125	0.084	0.020	0.115	0.074	0	0.083	0.061	0.036	0.063	0.032	0

Notes: See notes for table 3. Some values are repeated from table 3 for comparative purposes. Values that do not vary by clustering levels are not replicated in table 4. Clssrm: classroom; Tchr: teacher; Comm Arts: communication arts.

Finally, in an omitted analysis available from the authors upon request, we confirm our findings using a falsification exercise where we randomly assign teachers to TPPs. Given that table 3 shows nearly all of the variance in the TPP-effect estimates is driven by estimation error, we would expect to obtain similarly sized TPP "effects" even when we estimate models using randomly generated false TPP assignments. This is indeed what we find. We also show that improperly clustered models indicate non-zero TPP "effects" even when we randomly generate TPP assignments for teachers.[22]

6. ROBUSTNESS AND SENSITIVITY ANALYSIS

In this section we consider other factors that may explain our findings and examine the robustness of our results to an alternative methodological approach for isolating the true variance of the TPP effects.

Selectivity and Bias

Selective attrition and/or differential grade-level placements across TPPs that are correlated with teaching performance may bias our analysis toward finding no differences in effectiveness across teachers from different TPPs. Recall that the student–teacher linked data panel in Missouri does not have links prior to 2008–09, but in our analysis we use teachers who began teaching as early as the 2004–05 school year. Selective attrition will be an issue if, for example, ineffective teachers are both more likely to exit and also disproportionately come from some programs. If the ineffective teachers exit between their start dates and when we gain access to student–teacher links, then their effectiveness would not be captured by our models. This would shrink the observed variance across TPPs.

One way to directly address this issue would be to restrict our analysis to first-year teachers, for whom selective attrition is not an issue. In unreported results we re-estimate our models after restricting our data set to include first-year teachers only. A limitation of this analysis is that our sample size is dramatically reduced, which further weakens statistical power.[23] Noting this caveat, when we restrict the analysis to first-year teachers we find no discernible differences in the estimated TPP effects.

22. Gansle et al. (2010) report that their TPP estimates are "generally consistent" (p. 21) with those from their previous reports. This seemingly contradicts our finding that the TPP estimates are almost entirely composed of estimation error. There are two possible explanations. First, differences across TPPs could actually be larger in Louisiana, although the point estimates provided by Gansle et al. do not strongly support this hypothesis. A more likely explanation is that the Louisiana reports use an overlapping sample of teachers from year to year.
23. This limitation is an important part of the story. As a practical matter, the data simply do not facilitate narrow subanalyses in the context of estimating TPP effects.

Next we incorporate a related selection issue—that ineffective teachers may be less likely to teach in tested grades and subjects. Fuller and Ladd (2013) show that weaker teachers are more likely to teach in lower grades and that accountability pressure has exacerbated this trend. If this is the case in Missouri, and if the weak teachers who are kept out of tested grades and subjects are disproportionately trained at some TPPs, it would lead to an understatement of the variance of TPP effects (because the weakest TPPs would not look as weak as they should in our models). We incorporate this dimension of potential selectivity into our analysis by calculating "representation ratios" for TPPs. The numerator in these ratios includes the teachers in our analytic sample. The denominator includes teachers who were observed teaching in a self-contained Missouri elementary school classroom in any grade at any time after we began tracking certifications. That is, the denominator includes three types of teachers (or any combination therein): (1) teachers in our analytic sample, (2) teachers who teach in an elementary grade below grade 4, and (3) teachers who teach in a Missouri elementary school after we begin tracking certifications but quit before we obtain access to student–teacher links.

We perform a simulation exercise to determine whether the observed variance in the representation ratios across programs is more than we would expect to see by chance. Specifically, we randomly assign teachers across programs, holding program-level sample sizes fixed, and recompute the ratios based on the random assignments. We repeat this process 200 times to produce empirical confidence intervals for the cross-program variance of the ratios under random assignment. We compare the observed variance of the ratios across TPPs to the random assignment confidence intervals.

We cannot reject the null hypothesis that the representation ratios are constant across programs, although the variance of the observed ratios is on the high end of what we generate using the random-assignment simulations. Specifically, for the 24 main programs, the observed cross-program variance of the representation ratios is 0.0021, and the 95 percent confidence interval is 0.0006 to 0.0023. For the 12 large programs the observed variance is 0.0012 with a confidence interval of 0.0002 to 0.0013.

We also construct ratios to isolate variation in representation due to grade placement only (i.e., ignoring attrition). We obtain similar results for the full 24-program sample, but fail to reject the null for the large program subsample (p-value $\approx$ 0.03). One explanation for this result is selective grade placement as discussed by Fuller and Ladd (2013). Another is that there are small differences in grade-level emphasis across TPPs that are unrelated to quality.[24]

24. This emphasis could be formal or informal. For example, it could be that faculty at some institutions encourage/motivate teachers-in-training to teach in early or late elementary grades. Small

Overall, to the extent that we can examine the issue of selection into the analytic sample, there is little indication that systematic differences across TPPs in terms of teacher representation are biasing our results. Although we find no strong evidence to suggest that attrition and grade placement issues are driving our findings, it is worth noting that these issues will be relevant in any state's attempt to produce TPP rankings.

Sensitivity of Findings to Methodological Adjustments

The adjustment that we use to correct for estimation-error variance in the TPP coefficients is in the same spirit as similar adjustments that have been used in numerous prior studies (e.g., see Aaronson, Barrow, and Sander 2007; Koedel 2009; Mas and Moretti 2009; Rothstein 2010; Koedel and Betts 2011), but there are alternative approaches. To examine the robustness of our findings to one such alternative, we estimate a series of models that specify the teacher effects as random instead of fixed. The random-effects estimates are shrunken to account for statistical imprecision by weighting the TPP effect estimates toward the grand mean (or prior), with the weight on the grand mean increasing with statistical imprecision.[25]

Mirroring the analysis thus far, we compare output from models that differ by the clustering structure, which we impose using hierarchical linear models (Raudenbush and Bryk 2002). Table 5 shows the standard deviations of the shrunken TPP effect estimates, along with *p*-values from likelihood ratio tests for the joint statistical significance of the TPP random effects.[26] We estimate models analogous to models A and B from before. One issue with the random effects specifications is that the models partial out variance related to the control variables prior to estimating the TPP effects. So, for example, if teachers sort to schools along observable dimensions in ways related to quality, and this sorting is also correlated with TPP of attendance, it will put downward pressure on the variance of the TPP random-effect estimates. The similarity of results between models A and B, as shown in table 2, suggests that this is not an important problem in our data. Still, to assuage any lingering concerns, table 5 also shows estimates from a "bare" specification (labeled "Model A0"), which includes only students' lagged test scores and grade and year indicators. We interpret the TPP variance estimates from Model A0 as upper bounds.

differences across TPPs along this dimension could account for the variability across TPPs in the grade-placement ratios.

25. Of course, shrinkage estimators can be produced via other means as well. For example, the shrunken coefficients from Model B shown in figure 1 are produced using an ex post procedure as described in Koedel, Leatherman, and Parsons (2012).

26. We do not make any additional adjustments to the standard deviation estimates reported in table 5—we simply report the standard deviation of the shrunken estimates. The estimation-error adjustment occurs implicitly through shrinkage.

Table 5. Standard Deviations of Shrunken TPP Effects Using Random Effects Specifications with Different Levels of Clustering

	Mathematics			Communication Arts		
	Model A0	**Model A**	**Model B**	**Model A0**	**Model A**	**Model B**
24 Major Preparation Programs						
No additional random effects	0.035	0.034	0.036	0.032	0.029	0.029
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Classroom-level random effects	0.022	0.020	0.021	0.024	0.020	0.019
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Teacher-level random effects	0.014	0.009	0.008	0.020	0.014	0.012
	(0.03)	(0.21)	(0.32)	(0.00)	(0.02)	(0.08)
Programs Producing 50 or More New Teachers						
No additional random effects	0.032	0.029	0.028	0.025	0.020	0.016
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Classroom-level random effects	0.025	0.021	0.020	0.021	0.015	0.010
	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.09)
Teacher-level random effects	0.018	0.011	0.007	0.018	0.011	0.005
	(0.03)	(0.22)	(0.45)	(0.01)	(0.10)	(0.49)

Note: p-values from likelihood ratio tests for the joint statistical significance of the TPP effects are in parentheses.

The results in table 5 corroborate the findings from our main analysis. The most important differences are across models that differ by the clustering structure. Models that do not cluster at all, or cluster at only the classroom level, indicate much larger differences across TPPs than do models that cluster the data at the teacher level. The estimates in table 5 suggest that the real differences in performance across teachers from different TPPs are similar to or smaller than the differences indicated by our primary estimates in table 3.

7. DISCUSSION AND CONCLUSION

We evaluate traditional teacher preparation programs in Missouri using empirical models similar to those used in previous research studies (Boyd et al. 2009; Goldhaber, Liddle, and Theobald 2013) and at least two ongoing statewide evaluations (in Louisiana and Tennessee). The work we perform here is along the lines of what has been encouraged by the United States Department of Education (2011) and researchers from the Center for American Progress (Crowe 2010) and Education Sector (Aldeman et al. 2011), among others. Moreover, all twelve phase-1 and phase-2 Race to the Top winners have committed to using achievement data for public disclosure of the effectiveness of TPP graduates, and five winners have committed to using teacher effects on student

achievement for program accountability (Crowe 2011). A key finding from our study is that the measureable differences in effectiveness across teachers from different preparation programs are much smaller than has been previously suggested. This becomes apparent when we cluster the data to account for the fact that students who are taught by the same teacher do not represent independent observations by which the teacher's preparation program can be judged. We encourage policy makers to think carefully about our findings as achievement-based evaluation systems, and associated accountability consequences, are being developed for TPPs.

Our study also adds to the body of evidence showing that it is difficult to identify which teachers will be most effective based on pre-entry characteristics, including TPP of attendance. That said, work by Boyd et al. (2009) suggests that variation within TPPs in terms of preparation experiences may be large. For example, Boyd et al. find that better oversight of student teaching for prospective teachers is positively associated with success in the classroom later on. The current research literature is too thin to fully understand how differences in within-program experiences affect teaching performance, but it would not be unreasonable, for example, to expect that within-program variability in the quality of training exceeds across-program variability (in the student-teaching oversight example, this would be the case if the quality of prospective teachers' mentors is unrelated to TPP of attendance).

We conclude by noting that our findings need not be interpreted to suggest that formal, outcome-based evaluations of TPPs should be abandoned. In fact, the lack of variability in TPP effects could partly reflect a general lack of innovation at TPPs, which is facilitated by the absence of a formal evaluation mechanism. The mere presence of an evaluation system, even if it is not immediately fruitful, may induce improvements in teacher preparation that could improve in meaningful ways students' short-term and long-term outcomes (Hanushek and Rivkin 2010; Chetty, Friedman, and Rockoff 2014; Hanushek 2011). Still, we caution researchers and policy makers against overstating the present differences in TPP effects as statewide rankings become increasingly available. If administrators do not understand how small the differences in TPP effects really are, they could make poor hiring decisions by overweighting TPP rankings in their decisions.

The authors are at the University of Missouri, Columbia, in the department of economics. Koedel is also in the Truman School of Public Affairs. They thank Yi Du for valuable research assistance. They also gratefully acknowledge research support from the Center for Analysis of Longitudinal Data in Education Research (CALDER) and a collaborative relationship with the Missouri Department of Elementary and Secondary Education. The usual disclaimers apply.

REFERENCES

Aaronson, Daniel, Lisa Barrow, and William Sander. 2007. Teachers and student achievement in the Chicago Public High Schools. *Journal of Labor Economics* 25(1):95–135. doi:10.1086/508733

Aldeman, Chad, Kevin Carey, Erin Dillon, Ben Miller, and Elena Silva. 2011. A measured approach to improving teacher preparation. *Education Sector Policy Brief*.

Bertrand, Marriane, Esther Duflo, and Sendhil Mullainathan. 2004. How much should we trust differences-in-differences estimates? *Quarterly Journal of Economics* 119(1):249–275. doi:10.1162/003355304772839588

Bester, Alan C., Timothy G. Conley, and Christian B. Hansen. 2011. Inference with dependent data using cluster covariance estimators. *Journal of Econometrics* 165(2):137–151. doi:10.1016/j.jeconom.2011.01.007

Bill and Melinda Gates Foundation. 2013. *Ensuring fair and reliable measures of effective teaching: Cumulating findings from the MET Project's three-year study*. Seattle, WA: Bill and Melinda Gates Foundation.

Boyd, Donald, Pam Grossman, Hamilton Lankford, Susanna Loeb, and Jim Wyckoff. 2009. Teacher preparation and student achievement. *Educational Evaluation and Policy Analysis* 31(4):416–440. doi:10.3102/0162373709353129

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014. Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. *American Economic Review* 104(9):2633–79.

Clotfelter, Charles T., Helen F. Ladd, and Jacob L. Vigdor. 2006. Teacher-student matching and the assessment of teacher effectiveness. *Journal of Human Resources* 41(4):778–820.

Crowe, Edward. 2010. *Measuring what matters: A stronger accountability model for teacher education*. Washington, DC: Center for American Progress.

Crowe, Edward. 2011. *Getting better at teacher preparation and state accountability: Strategies, innovations and challenges under the federal Race to the Top program*. Washington, DC: Center for American Progress.

Ehlert, Mark, Cory Koedel, Eric Parsons, and Michael Podgursky (forthcoming). Selecting growth measures for use in school evaluation systems: Should proportionality matter? *Educational Policy*.

Fuller, Sarah C., and Helen F. Ladd. 2013. School-based accountability and the distribution of teacher quality across grades in elementary schools. *Education Finance and Policy* 8(4):528–559. doi:10.1162/EDFP_a_00112

Gansle, Kristin A., George H. Noell, R. Maria Knox, and Michael J. Schafer. 2010. Value added assessment of teacher preparation in Louisiana: 2005–2006 to 2008–2009. Unpublished paper, Louisiana State University.

Gansle, Kristin A., George H. Noell, and Jeanne M. Burns. 2012. Do student achievement outcomes differ across teacher preparation programs? An analysis of teacher education in Louisiana. *Journal of Teacher Education* 63(5):304–317. doi:10.1177/0022487112439894

Goldhaber, Dan, and Michael Hansen. 2010. Using performance on the job to inform teacher tenure decisions. *American Economic Review (P&P)* 100(2):250–255.

Goldhaber, Dan, and Roddy Theobald. 2013. Managing the teacher workforce in austere times: The implications of teacher layoffs. *Education Finance and Policy* 8(4):494–527. doi:10.1162/EDFP_a_00111

Goldhaber, Dan, Stephanie Liddle, and Roddy Theobald. 2013. The gateway to the profession: Assessing teacher preparation programs based on student achievement. *Economics of Education Review* 34(June):29–44. doi:10.1016/j.econedurev.2013.01.011

Hanushek, Eric A. 2011. The economic value of higher teacher quality. *Economics of Education Review* 30(3):466–479. doi:10.1016/j.econedurev.2010.12.006

Hanushek, Eric A. 2003. The failure of input-based schooling policies. *Economic Journal* 113(485):F64–F98. doi:10.1111/1468-0297.00099

Hanushek, Eric A., and Steven G. Rivkin. 2010. Generalizations about using value-added measures of teacher quality. *American Economic Review (P&P)* 100(2):267–271.

Harris, Douglas N., and Tim R. Sass. 2011. Teacher training, teacher quality and student achievement. *Journal of Public Economics* 95(7–8):798–812. doi:10.1016/j.jpubeco.2010.11.009

Koedel, Cory. 2009. An empirical analysis of teacher spillover effects in secondary school. *Economics of Education Review* 28(6):682–692. doi:10.1016/j.econedurev.2009.02.003

Koedel, Cory, and Julian R. Betts. 2007. Re-examining the role of teacher quality in the educational production function. University of Missouri Working Paper No. 07–08.

Koedel, Cory, and Julian R. Betts. 2011. Does student sorting invalidate value-added models of teacher effectiveness? An extended analysis of the Rothstein critique. *Education Finance and Policy* 6(1):18–42. doi:10.1162/EDFP_a_00027

Koedel, Cory, Rebecca Leatherman, and Eric Parsons. 2012. Test measurement error and inference from value-added models. *B.E. Journal of Economic Analysis & Policy* 12(1):1–37. doi:10.1515/1935-1682.3314

Levine, Arthur. 2006. *Educating school teachers*. Washington, DC: The Education Schools Project.

Lincove, Jane A., Cynthia Osborne, Nick Mills, and Amanda Dillon. 2013. The politics and statistics of value-added modeling for accountability of teacher preparation programs. *Journal of Teacher Education* 65(1):24–38. doi:10.1177/0022487113504108

Mas, Alexandre, and Enrico Moretti. 2009. Peers at work. *American Economic Review* 99(1):112–145. doi:10.1257/aer.99.1.112

Mihaly, Kata, Daniel McCaffrey, Tim R. Sass, and J. R. Lockwood. 2013a. Where you come from or where you go? Distinguishing between school quality and the effectiveness of teacher preparation program graduates. *Education Finance and Policy* 8(4):459–493. doi:10.1162/EDFP_a_00110

Mihaly, Kata, Daniel McCaffrey, Douglas O. Staiger, and J. R. Lockwood. 2013b. *A composite estimator of effective teaching*. Available www.metproject.org/downloads/MET_Composite_Estimator_of_Effective_Teaching_Research_Paper.pdf. Accessed 14 May 2014.

Noell, George H., Bethany A. Porter, and R. Maria Patt. 2007. Value added assessment of teacher preparation in Louisiana: 2004–2006. Unpublished paper, Louisiana State University.

Noell, George H., Bethany A. Porter, R. Maria Patt, and Amanda Dahir. 2008. Value added assessment of teacher preparation in Louisiana: 2004–2005 to 2006–2007. Unpublished paper, Louisiana State University.

Papay, John P., and Mathew A. Kraft. 2010. Do teachers continue to improve with experience? Evidence of long-term career growth in the teacher labor market. Unpublished paper, Harvard University.

Raudenbush, Stephen W., and Anthony S. Bryk. 2002. *Hierarchical linear models: Applications and data analysis methods*, 2nd ed. Thousand Oaks, CA: Sage Publications Inc.

Rothstein, Jesse. 2010. Teacher quality in educational production: Tracking, decay, and student achievement. *Quarterly Journal of Economics* 125(1):175–214. doi:10.1162/qjec.2010.125.1.175

Taylor, Eric S., and John H. Tyler. 2012. The effect of evaluation on teacher performance. *American Economic Review* 102(7):3628–3651. doi:10.1257/aer.102.7.3628

Tennessee Higher Education Commission. 2010. *Report card on the effectiveness of teacher training programs*. Nashville, TN: Tennessee Higher Education Commission.

United States Department of Education (USDOE). 2011. *Our future, our teachers: The Obama Administration's plan for teacher education reform and improvement*. Washington, DC: United States Department of Education.

Weisberg, Daniel, Susan Sexton, Jennifer Mulhern, and David Keeling. 2009. *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness*. Brooklyn, NY: The New Teacher Project.

Wiswall, Matthew. 2013. The dynamics of teacher quality. *Journal of Public Economics* 100(1):61–78. doi:10.1016/j.jpubeco.2013.01.006

APPENDIX A: SUPPLEMENTARY TABLE

Table A.1. Teacher Counts for Teacher Preparation Programs in Missouri That Produced More Than 15 Teachers in Our Final Analytic Sample

Program	New Elementary Teacher Count	First-Year Teacher Count
Program 1	143	29
Program 2	120	32
Program 3	118	40

Table A.1. Continued.

Program	New Elementary Teacher Count	First-Year Teacher Count
Program 4	111	22
Program 5	106	22
Program 6	76	18
Program 7	61	15
Program 8	57	11
Program 9	53	9
Program 10	53	8
Program 11	53	16
Program 12[a]	49	15
Program 13	39	6
Program 14	39	6
Program 15	34	7
Program 16	30	6
Program 17	29	6
Program 18	26	5
Program 19	24	6
Program 20	20	5
Program 21	19	8
Program 22	17	2
Program 23	16	5
Program 24	16	6

Notes: The teacher counts are from the analytic data sample.
[a]We include program 12 in our "large" program sample, although our findings are not qualitatively sensitive to dropping it from this group.

APPENDIX B: SELECTION INTO TPPS

Our study reveals very little variation in teacher quality across teachers who attended different TPPs. The fact that the differences across TPPs are small is surprising for two reasons. First is the presence of seemingly large input-based differences across teacher preparation programs (Levine 2006; Boyd et al. 2009). Our results, however, are broadly consistent with other research showing teaching effectiveness can rarely be linked to any observable characteristic of teachers (e.g., see Hanushek 2003; Koedel and Betts 2007). Second is that our estimates also embody differential selection into TPPs. Even if differences in inputs across TPPs are small, one might still expect differences in teacher performance across TPPs owing to differences in selection alone.

In table B.1 we briefly extend our analysis to examine the selection issue. The table uses supplementary ACT-score data from all college graduates at the eleven public-university TPPs in Missouri that are represented in our

Table B.1. Average ACT Scores by University for Eleven Public Universities Included in Our Evaluation

	Average ACT Scores		
	All Graduates	**Graduates with Education Major**	**Observed Elem Teachers**
University of Missouri-Columbia[a]	26.3	25.7	24.2
University of Missouri-Kansas City	25.3	23.8	22.8
Missouri State University[a]	24.7	24.1	21.7
Missouri Southern State University[a]	23.9	24.0	21.1
University of Missouri-St. Louis[a]	23.7	23.0	22.0
Southeast Missouri State University[a]	22.9	23.2	21.9
Northwest Missouri State University[a]	22.6	22.7	22.8
University of Central Missouri[a]	22.6	22.5	20.8
Missouri Western State University[a]	22.5	23.3	21.0
Lincoln University	21.4	22.0	21.0
Harris-Stowe State University	19.3	19.8	18.8
ACT variance across universities	3.68	2.17	1.93
ACT range across universities	7.0	5.9	5.4

Notes: The calculations in columns 1 and 2 are based on graduates from the listed universities who entered the public school system between 1996 and 2001 and graduated prior to 2009. The calculations in column 3 are based on data from the teachers we evaluate in our study. The population standard deviation in ACT scores, nationally, is approximately 4.
[a]Indicates the program is one of the twelve large programs in the state.

study.[27] The first column of the table shows the average ACT score for all graduates from each university. The second column shows average ACT scores for the subset of graduates who earn an education degree.[28] The third column shows average ACT scores for the graduates who actually end up working as elementary teachers and appear in our data.

The bottom rows of table B.1 reveal an interesting pattern: Differences in selectivity across institutions on the whole, as measured by ACT scores, are only partly reflected in the teacher population. Put differently, teachers from colleges that are differentially selective are more similar to each other than are typical students from the same colleges. Although table B.1 is merely descriptive and measures selection along just a single dimension, it offers some insight as to why selection may not be playing an important role in differentiating TPPs.

27. The higher education data come from cohorts of graduates who began their college careers between the years of 1996 and 2001 and completed their degrees at one of the specified universities prior to 2009. These data do not perfectly overlap with the cohorts of teachers we evaluate but should provide a fair representation.
28. Although not all of the teachers in our analysis are education majors, many are.